AD-A054 852 MARYLAND UNIV COLLEGE PARK DEPT OF COMPUTER SCIENCE F/G 6/4
SOME CURRENT CONCEPTS AND PROBLEMS IN PATTERN CLASSIFICATION AN--ETC(U)
APR 78 L N KANAL AFOSR-76-2901
UNCLASSIFIED TR-640 AFOSR-TR-78-0958 NL

| OF |
AD
A054852

END
DATE
FILMED
7-78
DDC

COMPUTER SCIENCE
TECHNICAL REPORT SERIES

UNIVERSITY OF MARYLAND
COLLEGE PARK, MARYLAND
20742

(2)

(14) Technical Report TR-64Ø (11) April 1978 (12) 37p.

ENG-73-Ø4099-A02

(15) ✓AFOSR-76-2901, ✓NSF-ENG-73-Ø4099

(6)

SOME CURRENT CONCEPTS AND PROBLEMS IN PATTERN CLASSIFICATION AND FEATURE EXTRACTION,

(10) Laveen N. Kanal

Laboratory for Pattern Analysis
University of Maryland
College Park, MD 20742

(9) Technical rept.,

(16) 23Ø4 (18) AFOSR

(17) A2 (19) TR-78-Ø958

Paper to appear in the Proceedings of the 41st Meeting of the International Statistical Institute.

409 022 LB

SOME CURRENT CONCEPTS AND PROBLEMS
IN PATTERN CLASSIFICATION AND FEATURE EXTRACTION

Laveen N. Kanal

University of Maryland
College Park, Maryland 20742
U.S.A.

1. INTRODUCTION

In [Kanal (1974)], while commenting on feature-subset selection, it was mentioned that the possibility of posing many problems in pattern recognition as graph-searching problems, suggested approaches likely to receive attention in the near future. The paper also mentioned the prospect of increasing interplay between pattern recognition and "problem-solving" techniques of Artificial Intelligence (A.I.) [Nilsson, (1971)].

The embedding of the statistical and syntactic models of pattern recognition theory into the state-space and AND/OR graphs and search strategies of A.I. and the development of connections between these various representations have, as anticipated led to new results and a somewhat different way of thinking about pattern recognition. For example, it has led us to view many pattern recognition problems in terms of a joint search of paths in a model space and a data space with feedback between the two searches.

Section 2 of this paper states some major limitations of the standard multivariate statistical approach to pattern classification and motivates the renewed interest in multiclass multistage classification. Section 3 categorizes recent contributions in the literature on multistage decision schemes. Section 4 outlines general state space representations and ordered search strategies, and section 5 illustrates how these serve as theoretical models for multistage multiclass pattern classification. General state-space graphs for nearest neighbor (NN) classification are described in section 6, which also mentions recent results on NN error estimation and bounds using labelled and unlabelled samples.

The shortcomings of the usual linguistic-syntactic models are listed in section 7 and used to motivate the problem-reduction representations defined in section 8, which comments on efficient feature extraction using problem-reduction and state-space representations. Section 9 contains concluding remarks on this and other research directions and problems. The bibliography guides the reader to recent papers, surveys, conference proceedings and edited collections covering broad areas of pattern recognition methodologies and their applications.

2. MULTISTAGE CLASSIFICATION SCHEMES: RATIONALE FOR RENEWED INTEREST

In various applications of current interest, e.g., biomedical image recognition and remote sensing, the number of features N is quite large and the classes M are multimodal and also numerous. For such problems multistage classification schemes are more frequently being used than "one-shot" classifiers which give an M way decision in a single step, and use a common set of features for all classes.

A straight casting of such pattern recognition problems into the standard molds of multivariate statistical classification theory usually requires the estimation of high dimensional unknown distributions and unknown a priori probabilities for many categories. Despite great theoretical attention, the practical modeling and estimation of highly multivariate distributions remains essentially infeasible.

The rationale for multistage classification is that decomposing the multiclass classification problem into several stages may simplify the decision making process in practice. For example, a priori (structural) knowledge concerning the physical and biological relationships between categories and groups of categories may be explicitly used to structure the skeleton of the decision tree.

If for an M-class, N-feature problem, processing has to be limited to no more than K<<N dimensional distributions, then different subsets of K features at the various stages promise better results than a single, "best" set of K features used in an M-way "one-shot" decision. As shown in [Cover and Van Campenhout (1976)] no non-exhaustive sequential K-feature selection procedures in which the subsets are constrained to nest, can be optimal even for jointly normal measurements.

Recall that to avoid exhaustive search, suboptimal procedures such as forward sequential, backward sequential and K-individually best features have been used together with various distance measures or other criteria [see Kanal (1974)]. Cover and Van Campenhout (1976) show that there exist *reasonable class p.d.f.'s for which these* search strategies lead to the worst possible K-feature subset, $1 < K < N$. For the normal model with $N = 4$ they give a simple numerical example in which the best 2-feature set consists of the two individually worst measurements and neither of the best 3-element or 2-element subsets includes the best subset of lower cardinality. In the context of hierarchical classifiers, it is easily proven [Kulkarni (1976)] that at each node of a given tree structure, the feature assignment which gives the node decision with the least average error, is not necessarily the overall optimal assignment of features. And in general, maximizing classification performance at each individual node of a decision tree does not result in an overall optimum decision tree.

Thus measuring a large number of common features on a large number of categories and then using the above mentioned, popular, feature subset selection procedures to reduce the feature set for subsequent design of a one-shot classifier, is unlikely to be satisfactory. Furthermore, even with multistage decomposition, straight optimizing of each node decision is unlikely to give good results.

To avoid the first pitfall, the original definition and extraction of features needs to be less arbitrary than is usually the case in multivariate statistical classification. Model-directed, data-confirmed, structural feature extraction seems relevant in many situations and is briefly described later in this paper. Also called for are systematic procedures for the general, global, design of hierarchical classifiers.

3. MULTISTAGE DECISION SCHEMES-A SELECTIVE SURVEY

Procedures proposed in the literature for designing multistage decision schemes consist of: (a) converting decision tables into optimal decision trees [e.g. Knuth (1971), Meisel & Michalopoulos tion rules; (c) hierarchical classification methods. In sequential classification schemes, the features are linearly ordered, but in most cases, no particular order is imposed on the class labels. Any class label may be accepted at any stage. Hierarchical classifiers are characterized by the hierarchical ordering imposed both on the features and the class labels. At each stage of the measurement process some classes are rejected from further consideration as candidates for the test sample's label.

The table conversion methods [Knuth (1971), Meisel & Michalopoulos (1973), Pollack et al (1971), Reinwald & Soland (1966), Winston (1969)] assume a decision table is given and do not address the problems of feature selection and classification error. For example in Meisel & Michalopoulos(1973) the feature space is already assumed to have been partitioned into the various decision regions. Then a recursive procedure arranges a set of piece wise constant boundaries in metric space so as to minimize the average number of comparisons needed to classify a sample. Thus the algorithm rearranges the order of the tests optimally without affecting the misclassification rate.

Most of the parametric and nonparametric multiclass classification schemes proposed in the literature may be described within the framework of state space graph models and ordered search strategies.

4. STATE SPACE REPRESENTATIONS AND ORDERED SEARCH

A state-space representation (SSR) for a problem consists of specifying a set $S = \{ s_i \}$ of state descriptions; a set $I \subseteq S$ of initial states; a set $S^* \subseteq S$ of goal states; state transition operators T which, applied to state s_i, produce successor states, $T(s_i)$. SSR is defined informally in [Nilsson (1971)] and formally in [Stockman (1977)] in which an ordered successor function $q: S \times N \rightarrow S$, with N the set of natural numbers, is used to define the successors of state s_i as $Q(s_i) = \{s_k : s_k = q(s_i, j)$ for some $j\}$.

A directed graph model for state-space search uses nodes of the graph to represent problem states and associates arcs of the graph with the operators. A state space representation is then viewed as an implicit definition of a directed graph and a search strategy becomes a process of "expanding" nodes step by step to obtain successor nodes, thereby making explicit a portion of the implicitly defined graph, in order to find a path from the initial node to a goal node.

Breadth-first, depth-first, best-first, and heuristic search are the descriptive names used in the (AI) literature [Winston (1977)] for some of the widely used search strategies, which are closely related to Dynamic Programming [Dreyfus & Law (1977)] Backtrack Programming [Golomb & Baumert (1965)] and Branch and Bound [Kohler & Steiglitz (1974)].

An ordered search strategy attempts to make the search efficient by ranking the nodes available for expansion at each state of the search according to some merit criteria. Assuming that a node has a finite number of successors, a general ordered search procedure is defined by the following algorithm.

Let $\bar{E}$ be the set of nodes which have been expanded (often called the CLOSED list), E the set of nodes which are candidates for expansion (the OPEN list) and T(s), the set of successors of node s. Initially set $\bar{E}$ to the empty set and E to the start node.

```
(1) If $E$ is empty exit with failure,
(2) Choose $s \in E$ such that $s$ has best merit, resolving
    ties arbitrarily,
(3) If $s$ is a goal node, exit with success obtaining
    the solution path by tracing back through the
    pointers,
(4) Remove $s$ from $E$ and place $s$ in $\bar{E}$. Expand node $s$
    generating all its successors; if there are no
    successors go to 1,
(5) For each successor $t \in T(s)$
    (a) if $t \notin \bar{E}$ and $t \notin E$, place $t$ in $E$ with a pointer
        to its parent node $s$ ; (b) if $t \in E$ or $t \in \bar{E}$ and
        new merit is better than old merit, place $t$ in $E$
        with pointer to $s$ and redefine merit of $t$,
(6) Go to 1.
```

The merit of a node may be defined via an evaluation function f(s) which uses selected features of a state s. A much used evaluation function is $f(s) = c(s) + h(s)$ where c(s) denotes the sum of the costs of the operators leading to the generation of state s from the initial state, and h(s) is a "heuristic" component, frequently based on some measure of difference between selected features of the state s and a goal state. Ordered search with this function is referred to as the A* algorithm [Nilsson (1971)]. In the directed graph representation c(s) is the cost incurred in going from the initial node to node s and h(s) is an estimate of the cost of a path from s to the nearest goal node. At a goal node s*, h(s*) is defined to be zero.

The above evaluation function may be generalized by defining $f(s) = (1-\alpha)c(s) + \alpha h(s)$, $\alpha \varepsilon [0,1]$. $\alpha = 0$ gives "uniform cost search", $\alpha = 1$ gives "pure heuristic search", while $\alpha = 1/2$ gives the previously defined "diagonal search" function used in the A* algorithm. Certain properties of A* were defined and investigated by Hart et al [1968]. Corrections [Hart et al (1972), Gelperin (1977)] and generalizations of certain types [Pohl (1970), Harris (1974), Martelli (1977)] have appeared subsequently.

A δ-graph has been defined as a graph with positive arc costs [Nilsson (1971)] and more generally [Vanderbrug (1977)], so as to allow finitely many arcs of zero cost. In either case, the following properties hold:

(a) an ordered search strategy is <u>complete</u> i.e., finds a solution whenever it exists, iff $\alpha \ \varepsilon \ [0,1)$;

(b) if a heuristic satisfies a lower bound condition $h(s) \leq h_\rho(d)$ for all nodes s, where h_ρ is the perfect heuristic (true remaining cost to nearest goal), and iff $\alpha \varepsilon [0,1/2]$ then search with $\mathring{h}$ is <u>admissible</u>, i.e., terminates with a minimum cost solution whenever one exists. Underestimating at each stage, the distance remaining to the goal, thereby underestimates the total path length. Since the actual cost along some completed non-minimum cost path cannot be less than an underestimate of the cost along an incomplete minimum cost path, the process of repeatedly extending the path which thus far has lowest underestimated cost will guarantee that the least cost path is found. Of course the closer h is to h_ρ the more efficient will be the search.

(c) to compare two admissible strategies using the ordered search algorithm, the concept of optimality of a search strategy is introduced. Note that admissibility refers to optimality of the solution. Let h_1 and h_2 be heuristic functions satisfying $h_2(s) < h_1(s) \leq h\ (s)$ for all non goal nodes s, and let $\alpha \varepsilon [0,1/2]$. Then for all δ-graphs containing a minimum cost solution, search with h_2 expands at least as many nodes as expanded by search with h_1. See also Gelperin [1977].

(d) If, for any two nodes s and s', which are connected by a path of cost c(s,s'), it is assumed that $h(s) - h(s') \leq c(s,s')$, then h is consistent. If h is consistent then for $\alpha = 1/2$, the ordered search algorithm never has to reopen a closed node, i.e., when it expands a node it has already found a minimum cost path to that node.

Vanderbrug (1977) substitutes easily followed geometric proofs for the algebraic proofs given previously for the above properties.

The next section illustrates how a generalization of the above approach serves as a theoretical model for multistage, multiclass classification [Kanal & Kulkarni (1976), Kulkarni (1976)].

5. STATE-SPACE GRAPHS FOR MULTICLASS CLASSIFICATION

In the state-space graph $G = \{S,E,F,W,c,r\}$ let a state $s \varepsilon S$ be a tuple $\{F_s, W_s\}$, where F_s is a subset of the total feature set F, and W_s is a subset of the total set of class labels W. W_s denotes the possible classifications that can be made on any path in the graph passing through the state s. An edge $e \varepsilon E$ represents the action of measuring a particular feature or set of features, and has an associated measurement cost determined by c, a non-negative real valued cost function. For a goal state s*, $F_s = \emptyset$ (null) and W_s contains one or zero (λ = reject) class labels. At a goal state, a misclassification risk r(s*), is incurred. The initial or start node of G contains all the possible class labels including λ, the reject class.

If N(s*) is the set of nodes on a path to a goal node s*, c(N(s*)) is the sum of arc costs along that path and r(s*) the risk at s*, then the total cost of making the decision s* is $f(s^*) = c(N(s^*)) + r(s^*)$. Two possible broad categories of classification schemes are:

(i) the risk r(s*) depends only on the features x_s measured along the path from the initial state to goal s*, and is denoted by

$r(s^*/x_s)$. An "S-admissible" strategy terminates at that goal s^* in G for which $f(s^*) = c(N(s^*)) + r(s^*/x_s)$ is minimum;

(ii) The risk $r(s^*)$ is a function of all the measurements, not just those on the path to s^*. If $x^{(k)}$ denotes the features observed until stage k, then the risk is $r(s^*/x^{(k)})$ and it could change as more features are observed until it reaches the value $r(s^*/x)$. A "B-admissible" strategy finds that (category) node s^* for which $f(s^*) = c(N(s^*)) + r(s^*/x)$ is minimum ; a Bayes optimal strategy results when arc costs are set to zero. In certain cases, B-admissible strategies to find the optimal category can be formulated without having to observe the total set of features.

Ordered search algorithms for S-admissible and B-admissible strategies for multiclass pattern classification can be realized by defining an evaluation function for node s as $f(s) = c(s) + h(s) + l(s)$, where $c(s)$ is the arc cost from the starting node to node s, $h(s)$ is an estimate of the arc cost from s to a goal node accessible from s, and $l(s)$ is an estimate of the risk of a goal node accessible from s.

S-Admissible Ordered Search Algorithm

This algorithm called Algorithm S, differs from algorithm A* described earlier, in the additional term used to estimate the risk at a goal. Let

$$l(s) \begin{cases} = r(s/x_s) \text{ if s if a goal node} \\ \leq \underset{j \in W_s}{\text{Min}} \left[\underset{y \in F(j) \sim F(s)}{\text{Min}} \quad r(j/x_s, Y) \right] \text{otherwise} \end{cases}$$

Here $F(j)$ denotes the measurement space spanned by the set of features measured on the path to node j, and $Y \in F(j) \sim F(s)$ is a vector in the complement space of $F(s)$ with respect to $F(j)$. Also

$$h(s) \begin{cases} = 0 \quad \text{if s is a goal node} \\ \leq \underset{j \in W_s}{\text{Min}} \quad c(s,j) \text{ otherwise} \end{cases}$$

where c(s,j) is the sum of arc costs from s to j. Closing a node implies observing a particular feature set associated with the node. An S-admissible strategy terminates when it first puts a goal node on the CLOSED list.

B-Admissible Ordered Search Algorithm

If the goal risk were to change with additional measurements taken on other paths, then the optimality of the first goal put into CLOSED cannot be guaranteed, because the additional observations may increase that goal's risk while decreasing the risk of some other goal. Algorithm B, which gives B-admissible search strategies uses an upper bounding function on the risk and a lower bounding function analogous to that used by Algorithm S.

Now the term l(s) which estimates the risk is defined by

$$l(s) \begin{cases} \leq \min\limits_{Y \epsilon F \sim F^{(k)}} r(s/x^{(k)}, Y) & \text{for a goal node} \\ \leq \min\limits_{j \epsilon W_n} \left[\min\limits_{Y \epsilon F \sim F^{(k)}} r(s/x^{(k)}, Y) \right] & \text{for s a non-goal node,} \end{cases}$$

where $x^{(k)}$ denotes the vector observations on a test sample after k stages of the algorithm and $Y \epsilon F \sim F^{(k)}$ denotes a random vector in the complement space with respect to the total feature set F. In algorithm B for each goal node s put on closed, an upper bounding function b(s) is computed, with $b(s) = c(s) + u(s/x^{(k)})$ and u satisties the inequality

$$u(s/x^{(k)}) \geq \max\limits_{Y \epsilon F \sim F^{(k)}} r(s/x^{(k)}, Y)$$

If s* is the goal node with minimum b(s), and if for all nodes in the union of OPEN and CLOSED, $b(s^*) \leq f(s)$, then the algorithm exits with s* as the B-optimal category. Algorithm B does not terminate when a goal node is put in the CLOSED list. Instead, at each iteration after a goal node is put on CLOSED, Algorithm B checks if there is some goal node on CLOSED such that the upper bound on its cost for any possible future measurement sequence, is less than the lower bound on the cost of any other goal node, either in CLOSED, or below some OPEN node in the graph. If so, the algorithm terminates with that goal node as the decision.

For certain parametric probability functions it may be possible to derive tight bounds on the risk of the goal accessible from a node s, as is shown in [Kulkarni & Kanal (1978)] where proofs for various properties of the algorithms appear. That paper also shows by example how B-admissible search of a state-space graph for leukocyte classification improves on the usual decision tree approach, in which, after reaching a terminal, no logical strategy exists for reconsidering classes possibly discarded earlier.

6. NEAREST NEIGHBOR CLASSIFICATION

For the nonparametric case, nearest neighbor (NN) classification rules are receiving increasing attention. See e.g.,[Dasarathy & Sheela (1977), Kanal (1974)]. Given an unlabelled (test) random sample x, the search for its nearest neighbor among a set of labelled (design) samples can be modelled as searching a state space graph, G(S,E,X,D,c,d). Here D refers to the total set of design samples Y_1, Y_2, ... Y_n, S is the set of states (nodes) and each state $s \in S$ is a tuple (F_s, D_s) where F_s, defined by the features measured on the path to s, is a subspace of the feature space F. D_s denotes a subset of D; for a goal state $|D_s| = 1$, i.e., it consists of a single labelled sample. At a goal node, d is the distance or other similarity measure between the test sample X and the design sample represented by the goal node. E and c refer, as before, to the edge set and cost function.

Depending on the cost functions defined for the edges and goal nodes various NN rules can be defined. For example, if c(s*) is the sum of arc costs on the path to goal node s*, Ys* is the labelled sample represented by s*, $d(X, Y_{s*} ; F_{s*})$ is the distance measured in the subspace $F_{s*} \subseteq F$, then one NN procedure [Kulkarni (1976)] defines the optimal goal s* to be the one for which f(s*) = $c(s*) + d(X, Y_{s*} ; F_{s*})$ has minimum value. Note that in this procedure, the distance is computed only in the subspace defined by the features measured in getting to s*. Such a procedure may be suitable when certain features are significant while others, are irrelevant to defining the similarity between a random test sample and a labelled sample from a particular class.

An alternative would be not to assume that some features are unimportant for the distance computation while retaining the tradeoff between feature measurement cost (arc costs) on the path to the goal, and the risk of not finding the true nearest neighbor. In this case, the optimal goal s* minimizes $f(s^*) = c(s^*) + d(X,Y_{s^*};F)$.

The conventional NN schemes are special cases of the above, obtained by setting arc costs to zero, i.e., now $f(s^*) = d(X,Y_{s^*};F)$. The above three general NN procedures can be implemented as state space searches giving S-admissible, B-admissible and B-admissible with zero arc costs, procedures respectively. Corresponding to the earlier problem of finding upper and lower bounds on the risk in the parametric cases, is the problem here of computing lower and upper bounds on the d( ) measure. Bounds for various metric and nonmetric similarity measures are derived in [Kulkarni (1976), Kulkarni & Kanal (1978)]. The extension to K-NN calculations is immediate.

A branch and bound algorithm for computing nearest neighbors by Fukunaga and Narendra (1975) is a special case of B-admissible search with zero arc costs. First the prototype samples are hierarchically decomposed into disjoint subsets, represented by a tree structure; any clustering technique can be used with computational efficiency, rather than meaningful groupings, being the main criterion. Two rules are used in the algorithm. Let S_p be the set of samples associated with node p in the tree. Let M_p be the mean of S_p and let γ_p be max $\{d(X_i, M_p) | X_i \varepsilon S_p \}$. Let B be the distance to X of the current nearest neighbor. Then Rule 1 in the Branch and Bound algorithm is: discard $X_i \varepsilon S_p$ as the potential nearest neighbor of X if $B + \gamma_p < d(X,M_p)$. Rule 2 is: Discard $X_i \varepsilon S_p$ if $B + d(X_i,M_p) < d(X,M_p)$. Several computational experiments are reported.

For example, the preprocessing step of dividing 1000 bivariate gaussian pseudorandom samples into 27 final subgroups, by successively applying the three-means algorithm, took 12,000 distance computation. When the branch and bound algorithm was applied to an additional 1000 samples, on average 61 distance calculations were required and no test sample took more than 87. For 3000 samples uniformly distributed in 8 dimensional space, the number of groups was 256, and on average a NN search took 451 distance calculations. Branch and Bound algorithms for feature subset selection and clustering appear in [Fukunaga and Narendra (1976)].

To aid K-NN computation, Friedman, Basket & Shustek (1975) sort the labelled samples on the values of one of the ℓ coordinates. For each test point, the prototypes n are examined in the order of the projected distance d_1 from the test point on the sorted coordinate. When $d_1 > d_\ell$ the ℓ-space distance to the Kth closest point of those already examined, no more prototypes need be examined. Best behavior is obtained by sorting on the values of each of the coordinates independently, and then selecting the axis with the smallest projected local density, i.e., the largest spread. Sparsity is calculated over a set of n_ℓ points centered on the test sample using for a value of n_ℓ the expected number of distance calculations under a uniform distribution.

Assuming a uniform distribution and assuming that sorting is performed on one coordinate at random, the expected number of distance calculations is

$$E[n_\ell] \leq \Pi^{-\frac{1}{2}} \, [K \, \ell \, \Gamma \, (\ell/2)]^{\frac{1}{\ell}} \, (2n)^{1-\frac{1}{\ell}}$$

Preprocessing is proportional to n ℓ log n.

Simulation experiments using the distributions: uniform on the unit square, bivariate normal, bivariate Cauchy showed that the analytical expression provides a close upper bound for actual average performance. For example, for $\ell = 2$, with $n = 1000$, NN search would require an (upper bound) average of 112 distance calculations.

The efficiency of this K-NN algorithm decreases slightly with K and more rapidly with the number of dimensions ℓ. For example for $\ell = 8$, and $n = 1000$, with $K = 1$, the number of distance calculations is upperbounded by 60% of the number for a brute force calculation. If N_t denotes the total number of test samples, a rough approximation of the breakeven point for using this procedure over the bruteforce method is $N_t \sim \frac{n \log n}{n-E(n_\ell)}$.

Kulkarni (1976) using examples from Euclidean measure, and similarity measures for binary vectors, showed that one can, relatively inexpensively obtain bounds which ordered search S-admissible and B-admissible algorithms to reduce the measurement cost or the number of distance computations needed to classify a test sample. Computational results similar to the branch and bound method, which is subsumed, could be anticipated. Analysis of the expected number of distance calculations remains to be done.

Recent theoretical results on K-NN performances bounds and error-reject estimation further motivate interest in NN search algorithms.

Given an unlabelled sample X whose label $\theta \epsilon [1,2...M]$ is to be decided, and given a finite set of n labeled samples (X_1,θ_1), $(X_2,\theta_2)...(X_n,\theta_n)$, a rule is called K-local if the decision $\hat{\theta}$ depends only on those pairs (X_i,θ_i) for which X_i is one of the K-NN of X. Let L_n denote the K-local rule error probability, and let $\hat{L}_n^R$, $\hat{L}_n^D$ and $\hat{L}_n^H$ denote respectively the resubstitution error estimate, the deleted ("leave-one-out") estimate and the hold out estimate. Rogers and Wagner (1978) prove that $E(L_n - \hat{L}_n^D)^2$ is bounded by A/n where A is an explicitly given small constant depending only on K and the number of categories M. The bound does not depend on the number of dimensions, ℓ, which suggests that local rules exchange K for ℓ [see Cover & Wagner (1976) on this and other topics in non-parametric discrimination, finite memory learning and pattern complexity]. Recently Devroye and Wagner (1977a,b) presented distribution-free bounds for

Prob $\{|\hat{L}_n - L_n|\} \geq \varepsilon\}$ where $\hat{L}_n$ stands for $\hat{L}_n^R$, $\hat{L}_n^D$ and $\hat{L}_n^L$.

A modified K-NN rule is obtained by allowing rejects. The modified rule, denoted as the (K,K') - NN rule makes the same decision as the K-NN rule whenever one or more labels receives at least K' votes from among the K-NN of X; otherwise the test sample is rejected. Let $E_{k,k'}$ and $R_{k,k'}$ denote the error and reject rates respectively, for the (K,K')-NN rule. Reject rates may be obtained from unlabelled samples. Devijver (1976) presents a distribution-free relationship between $E_{k,k'}$ and $R_{(k+1),(k+1)'}$ which allows the error rate to be obtained without having to label test samples and count errors. The price is that in addition to the K-NN's, the (k+1)st - NN will have to be found. The exchange is between the labels of the test samples and the label of the (k+1)st NN. Devijver's non-parametric results seems to improve the prospect for practical estimation of error rates from unlabelled samples. As noted in [Kanal(1974)] experience in the parametric case suggests that the error rate predicted from the emperical reject rate can be quite inaccurate if the model assumed in designing the classifier were inaccurate.

Another theoretical connection between K-NN's and error rates is developed by Tebbe (1976). He shows that the kth coefficient, in an orthogonal Legendre series expansion of the Bayes risk function for the two-class, zero-one loss case, can be estimated from an expectation defined in terms of the k+2-NN's of the patterns in a random sample.

7. STRUCTURAL FEATURE EXTRACTION, AND LIMITATIONS OF SYNTACTIC PATTERN RECOGNITION

Feature extraction techniques based on Fourier and other integral transforms, matrix methods and linear operator theory abound in pattern recognition theory. Non-linear feature extraction transformations are also being investigated. For example, in [Starks and de Figuieiredo (1977)] the transformation attempts to preserve graph theoretic structures such as minumum spanning tree, maximally complete subgraphs, inconsistant edges and diameter edges derived from data points.

For complex problems such as detecting and identifying structured elements in noisy biomedical waveforms or aerial photographs, most of these feature extraction methods provide a round-about, inefficient way of recapturing structures of interest apparent in portions of the scene. Future electro-optical and biologically motivated implementations may change this appraisal.

We would like to efficiently extract primitive structural elements (morphs) which are perceptually higher level objects than scalar measurements and use a variety of relationships among them in describing and recognizing patterns. Syntactic pattern recognition [see Fu (1974)] was presumably intended to overcome some of the limitations of statistical pattern recognition. However, syntactic pattern recognition has also viewed primitive extraction as preprocessing. This requires that all possibilities be considered in all regions of the data. Also separating the extraction of morphs from the analysis of structure excludes each process from information available to the other.

Formal language theory had addressed some problems of ambiguity and error. Earley's parser [Earley (1970)] was developed to handle ambiguous context-free grammars (CFG). Aho and Peterson (1972) showed how to model errors of insertion, deletion and mutuation of terminal symbols by adding productions to a grammar, and developed a minimum-distance error-correcting parser from the Earley parser. Lyons (1974) developed a least errors parser by extending the Earley parsing algorithm rather than the grammar. Assuming syntactic models for pattern noise and deformation are available (unlikely to be true for real data), Lu and Fu (1976) added probabilities to the techniques of Aho and Peterson (1972) and used Earley's algorithm to develop a maximum likelihood parse for any string over the terminal vocabulary.

While partially addressing ambiguity of analysis and description, syntactic pattern recognition has completely ignored ambiguous detection of primitives. If very small low level primitives are used, e.g., line or polynomial segments derived from a piecewise functional approximation of an entire waveform, ambiguity of detection may be avoided but other problems are created. The strings become too long-which makes parsing economy critical, and the segmentations are not anthropomorphic, which creates a need for grammatical inference.

Casting pattern analysis and description directly into the mold available from formal language theory, syntactic pattern recognition took on the burden of the concatenation relation and left-right parsing. Also, in general, a one-directional, i.e., strictly top-down or strictly bottom up parse procedure has been adopted. Strictly bottom-up methods [e.g. Fu (1974) Ledley (1966), Horowitz (1975), Pavlidis (1976)] do not take advantage of a priori knowledge during segmentation, while strictly top-down methods [e.g. Harlow and Eisenbeis (1973), Stockman, Kanal & Kyle (1976), Walker (1974)] can inefficiently generate hypotheses that are in no way related to a given instance of data.

Some desired objectives for a structural analysis procedure are: (1) the pattern analysis should be able to proceed in a bi-directional data-directed and model-directed manner with primitive extraction and structural analysis coupled; (2) the analysis should not be restricted to a cannonical left-right scan of the entire data but should be non-left-right, selective, and focus on prominent morphs; (3) multiple and ambiguous interpretations should be developed on a best-first basis, with ambiguity permitted in both segmentation and structural analysis.

For speech recognition, Miller (1973) and Reason (1976) have used context-free grammar models with non-left-right analysis, and they and others [Reddy (1973), Walker (1974)] allowed ambiguous detection of vocabulary terminals. A structural analysis paradigm which realizes the above enumerated objectives-not satisfactorily addressed previously-is developed in [Stockman (1977a)] which also presents the implementation of the paradigm in an extensively tested waveform parsing system. Recently the approach has also been used for the recognition of objects in imagery [Stockman (1977b)]. This non-directional approach begins by identifying certain prominent primitive components which can be reliably extracted and then uses model-directed, data-confirmed search for the remaining pattern structure. The theoretical concepts underlying the algorithm are mentioned next.

8. AND/OR GRAPHS (AG'S), STATE-SPACE REPRESENTATION (SSR) AND NON-DIRECTIONAL ANALYSIS (NDA) IN FEATURE EXTRACTION

A Problem-Reduction representation (PRR) recursively tries to solve a problem by transforming it into several simpler equivalents, any one of which if solved, solves the problem, or transforming it into several subproblems, all of which if solved, solve the original problem. Using nodes to represent problems and subproblems PRR's are modeled by AND/OR graphs (AG's) in which equivalent problems are represented by OR nodes and subproblems of a node are represented by AND nodes. The edges leading to the AND nodes are tied together with an arc to indicate that all of the AND descendents of a node must be solved in order to declare their parent solved. By inserting dummy OR nodes, mixed AND/OR nodes can be represented by combinations of pure OR nodes and pure AND nodes.

Solving a problem at the root node of an AG involves searching (making explicit) portions of the AG to find primitive problems whose solution allows the original problem to be declared solved, or showing that no such solvable primitive problems are present in the AG, in which case an "unsolvable" declaration is passed back up the AG. Solvable primitive problems are represented by terminal nodes in the AG. Costs of transforming problems into equivalent problems and subproblems may be associated with the edges of an AG.

An informal presentation of PRR and AG's is given in Nilsson (1971). Hall (1973) showed the equivalence of a CFG to a finite AG, and Chang and Slagle (1971) gave one approach to converting PRR to SSR so that the A* ordered search algorithm can be used to find optimal solution graphs in PRR according to a sum of edges cost criteria. Vanderbrug and Minker (1975) gave a formal treatment of PRR and showed an approach to bidirectionally relating AG search and state-space search. A different formal treatment and a different conversion between PRR and SSR is given in [Stockman (1977a)]. Being motivated by non-directional structural feature extraction this is the one of interest here.

A PRR is a 5 tuple $\{P,r,t,u,B\}$ where $P = \{P_i\}$ is an enumerable set of problem descriptions, $B \subseteq P$ is a set of initial problems only one of which need be solved, r is the ordered successor function $r: PxN \rightarrow P$, where N is the set of natural numbers, t is the node type function $t: P \rightarrow \{AND\ OR\}$ and u is the node solution function $u: P \rightarrow \{Live, Solved, Dead\}$. A problem is live when it is not known to be solved or dead, i.e., unsolvable. Solvability of OR nodes implies solvability of their parent, while unsolvability of any AND node implies unsolvability of its parent. Let $R(P_i)$ denote the set of all successors of problem P_i. A problem $i \varepsilon$ PRR is solved iff (1) $u(i) = solved$; or (2) $u(i) = Live$ and there exists successor $k \varepsilon R(i) \rightarrow u(k) = Solved$ and $t(k) = OR$; or (3) $u(i) = LIVE$, and for any successor $k \varepsilon R(i)$, $u(k) = SOLVED$ and $t(k) = AND$, PRR has a solution iff some problem $i \in B$ is solved.

Every OR successor of a problem (node) is called a primary successor but only the first AND successor of a node is a primary successor. There is no point in examining any AND alternative if a previous AND subproblem is unsolvable. Hence AND alternatives are considered sequentially and the primary successor is examined first. A primary descendent of the original problem (the root node) is either a primary successor of the root, or the primary successor of some primary descendent of the root.

Recognition of a solved problem (primary terminal) triggers the search, under the a priori constraints embodied in the PRR, for the solution to problems for which the solved problem is a primary successor. Typically, this would involve a top-down (model-directed) search for the solution of other non-primary successors. If the Inverse of the primary successor relation is available in the PRR, as is the case for CFG's and hence for finite AG's, the analysis can proceed recursively in either bottom-up or top-down direction.

For the feature extraction application primary terminal nodes represent the prominent morphs which can be reliably extracted from the data without any syntactic information. Problems which are not primary are always solved with respect to other morphs and properly related syntax. (The data segmentor is only asked to do work that is consistent with the global segmentations/interpretation being maintained by the structural analyzer.)

Individual morphs are defined as constrained mathematical curves and least-squares theory is the basis for morph detection and quality evaluation. When recognized, each substructure of the PRR must be assigned a quality $Q \leq 1.0$ to reflect the confidence of recognition. For primary terminals, Q is obtained from the morph primitive detector itself. For secondary morphs, Q depends not only on the quality of the detection but also on the degree to which the detection satisfies the structural hypothesis. For a non-primitive structure, the quality may be defined by the minimum quality of its substructures. The merit of a path in model-space is defined as the minimum quality of any structures identified along that path. The ordered search for interpretations will find the highest quality one first because it always extends the highest merit path first.

Multiple interpretations and non-directionality are facilitated by converting the PRR into an equivalent SSR such that partial solution trees in the AG, i.e., partial interpretations of the data, become encoded as states in SSR. Best-first search in this SSR can be shown to produce the minimax (best) solution tree (interpretation) in PRR. In practice, if the intervals of search for primitive features are not tightly constrained by syntax an inadmissible but efficient heuristic detection strategy is to scan exhaustively only once for morphs of certain minimum size, and then using pertubation operators in a state space search, grow each detection so long as quality is acceptable. In any event, as in the multiclass pattern classification work described earlier, the unifying concept between the structural analysis and detection algorithms is that of state-space search.

Because of the correspondence between CFG's and finite AG's, the new NDA algorithm in [Stockman (1977a)] can be viewed either as a "problem solver" or a parser. When applied to AG's representing games, i.e., game trees, the algorithm appears competitive with, although different from, the α-β tree pruning procedure [Nilsson (1971)] in efficiently producing the minimax solution tree. Knuth and Moore (1975) have procedurally defined the α-β method, analyzed its performance under some distributional assumptions and shown it to be optimal according to certain statistical criteria. Other analyses appear in [Fuller, Gaschnig and Gillogly (1973)] and [Newborn(1977)]. Apart from some simulations, no such analysis of the NDA algorithm has been done.

In AG's an underlying assumption is that subproblems can be solved independently. Levi and Sirovich (1976) defined Generalized AND/OR graphs (GAG's) in which subproblem interdependence is allowed. Such GAG's and other formulations of GAG's enlarge the potential representations for which Stockman's NDA algorithm and conversions to SSR may be defined and thereby enlarge the models available for structural representation and analysis.

9. MORE TO READ AND THINK ABOUT

The above tutorial presentation was designed to enable access to some of the literature and methodology behind some innovative approaches to multivariate statistical classification and structural feature extraction theory and practise, which are quite different from what currently appears in statistical journals. Much recent work in the statistical and engineering literature on pattern recognition is along lines already covered in Kanal (1974) .

The conference proceedings, books and edited collection, surveys and reports on prospects, listed in the bibliography ease the task of covering the proliferating literature on pattern recognition techniques and applications. [Agrawala (1976)] reprints two historically important out of print reports by E. Fix and J.L. Hodges, which motivated the later interest in NN methods. Far removed from the pattern recognition practitioners' present concerns are the fascinating, but rather difficult to follow works of Ulf Grenander (1976) on pattern synthesis, and William Hoffman (1976) on a Lie transformation group theory of form perception and feature extraction.

Of immediate concern are certain practical problem of measurement complexity and error estimation. Waller & Jain (1976), using a model [Abend et al (1965)] with first order nonstationary Markov dependent binary features, showed that independence of measurements is not a necessary condition either for the absence of the peaking phenomenon of measurement complexity [Chandrasekaran & Jain (1977), Van Ness (1977)] or for perfect discrimination. Van Campenhour (1977) resolves the paradox of the peaking phenomenon by showing that in a true Bayesian formulation it is attributable to improper comparisons of statistically incomparable models. In practise the phenomenon exists. In error estimation, Toussaint (1975), using a non parametric classifier, concluded that an estimator should be formed by equally weighting the resubstitution and rotation estimators while MacLachlan (1977), on the basis of asymptotic results for parametric classification using multivariate normal populations, suggests that very little weight should be assigned to the resubstitution estimator. Clearly specific examples may serve as counterexamples only but not as the basis of otherwise drawing general conclusions.

In addition to the problems cited earlier, much remains to be done on the decision-tree and state-space search formulation of hierarchical classifiers including the development of optimal procedures for continuous and mixed random variables. A list of general problems of automatic and semi-automatic pattern recognition appears in Kanal (1977) .

ACKNOWLEDGEMENTS

I am grateful to Prof. C.R. Rao for his invitation to write this paper for the 41st meeting of the International Statistical Institute. This research was supported by NSF grant ENG 73-04099-A02 and AFOSR grant 76-2901B.

BIBLIOGRAPHY

(1) Abend, K., Harley, T.J. and Kanal, L.N. (1965). Classification of binary random patterns. IEEE Trans. Inform. Theory, Vol. IT-11, 538-544, October 1965.

(2) Agrawala, Ashok K., ed. (1976). MACHINE RECOGNITION OF PATTERNS. IEEE Press Selected Reprint Series.

(3) Aho, A.V. and Peterson, T.G.(1972). A minimum distance error-correcting parser for context free languages. SIAM J. Comput., Vol. 4, December 1972.

(4) Bartels, P.H. and G.L. Wied (1977).Computer analysis and biomedical interpretation of microscopic images: current problems and future directions. Proceedings of the IEEE, Vol. 65, No. 2, 252-261, February 1977.

(4b) Batchelor, B., (ed), (1977), Pattern Recognition-Ideas in Practice, Plenum, 1977.

(5) Ben-Bassat, Moshe (1976). Multimembership classification with an application to medical diagnosis. IEEE Transactions on SMC, August

(6) Broffitt, James D., Ronald H. Randles and Robert V. Hogg (1976). Distribution-free partial discriminant analysis. Journal of the American Statistical Association. Vol. 71, No. 356, 934-939.

(7) van Campenhout, Jan M. (1977). On the peaking of the mean recognition accuracy: the resolution of an apparent paradox. Stanford University.

(8) Chan, Linda A., June Aono Gilman and Olive Jean Dunn (1976). Alternative approaches to missing values in discriminant analysis. Journal of the American Statis. Assoc. Vol. 71, No. 356, 842-844.

(9) Chandrasekaran, B., A.K. Jain (1977).Independence, measurement complexity and classification performance-an emendation. IEEE Transactions on Systems, Man & Cybernetics, Vol SMC - 7, No. 7, 564-566.

(10) Chang, P.C. and A.A. Afifi (1974).Classification based on Dichotomous and continuous variables. Journal of the American Statis. Assoc. Vol. 69, Number 351, 336-339.

(11) Chang, C.L. and J.R. Slagle(1971). An admissible and optimal algorithm for searching AND/OR graphs. Artificial Intelligence, Vol. 2, 117-128.

(12) Chen, Zen and King-Sun Fu (1977). Nonparametric bayes risk estimation for pattern classification. IEEE Transactions on SMC, Vol. SMC-7, No. 9, 651-656.

(13) Cover, Thomas, M., and Aaron Shenhar (1977). Compound bayes predictors for sequences with apparent Markov structure. IEEE Trans. on SMC, Vol. SMC-7, 421-424.

(14) Dasarathy, B.V., Sheela, B.V. (1977). Visiting nearest neighbors-a survey of nn classification techniques. Proc. 1977 International Conference on Cybernetics & Society, 630-636.

(15) Davies, H.I. and Edward J. Wegman (1975). Sequential nonparametric density estimation. IEEE Trans. on Information Theory, IT-21, No. 6, 619-628.

(16) Devijver, P.A., (1976). Error reject relationships in nearest neighbor decision rules. Proceedings 3rd International Joint Conference on Pattern Recognition, 255-259.

(17) Devroye, L.P. and T.J. Wagner. Distribution-free performance bounds with the resubstitution error estimate. University of Texas, Electrical Engineering Department, Austin, Texas.

(18) Devroye, L.P. and T.J. Wagner (1976). Nonparametric discrimination and density estimation. University of Texas at Austin, Information Systems Research Lab, Technical Report no. 183.

(19) Diday, E. and J.C. Simon (1976). Clustering Analysis. Comm. and Cybernetics 10, Digital Pattern Recognition, K.S. Fu, Ed. Springer, 47-94.

(20) Dreyfus, S.E. & A.V. Law (1977). The Art and Theory of Dynamic Programming. Academic Press.

(21) Dubes, Richard and Anil K. Jain (1976). Clustering Techniques: the user's dilemma. Pattern Recognition, Vol. 8, 247-260.

(22) Earley, J.C. (1970). An efficient context-free parsing algorithm. CACM, 12, 2, 94-102.

(23) Erickson, J.D. (1975). Advances in Automatic extraction of earth resources information from multispectral data. Proc. First Earth Resources Survey Symposium, Houston, Texas, 1-B, 1245-1274.

(24) Friedman, J. (1975). A variable metric decision rule for nonparametric classification. SLAC-PUB-1573, Stanford Linear Acceleration Center, Stanford.

(25) Friedman, J.H., Basket, F., and Shustek, L.J., (1975). An algorithm for finding nearest neighbors. IEEE Transactions on Computers, C-24, 750-753.

(26) Fritz, Jozsef (1975). Distribution-free exponential error bound for nearest neighbor pattern classification. IEEE Transactions on Information Theory, Vol. IT-21, No. 5, 552-557.

(27) Fu, K.S. (1968). Sequential Methods in Pattern Recognition and Machine Learning, Academic Press.

(28) Fu, K.S., (1976). Digital Pattern Recognition, Communications and Cybernetics. Vol. 10, Springer-Verlag.

(29) Fu, K.S., (1976). Pattern recognition in remote sensing of the earth's resources. IEEE Trans. on Geoscience Electronics, Vol. GE-14, No. 1, 10-18.

(30) Fu, K.S. (ed)(1976). Special issue on "Pattern Recognition Computer, Vol. 9, No. 5.

(31) Fu, K.S. (ed)(1977). Syntactic Pattern Recognition Applications, Springer Verlag.

(32) Fu, K.S., and T.L. Booth (1975). Grammatical Inference: Introduction and Survey. IEEE Transactions SMC, Vol. SMC-5.

(33) Fu, King-Sun and Azriel Rosenfeld (1976). Pattern Recognition and image processing. IEEE Transactions on Computers, Vol. C-25, No. 12.

(34) Fukunaga, Keinosuke and Patrenahalli M. Narendra (1976). Combinatorial problems in pattern recognition. Purdue University, School of Electrical Engineering, Report TR-EE 76-43.

(35) Kukunaga, K., Narendra, P.M. (1974). A branch and bound algorithm for computing k-nearest neighbors. IEEE Trans. Comp. Vol. C-24, No. 7.

(36) Fuller, S.H., J.G. Gaschnig, & JJ. Gillogly (1973). Analysis of the alpha-beta pruning algorithm. Department of Computer Science, Carnegie-Mellon University, Pittsburg, PA.

(37) Gardner, M.J. & D.J.P. Barker (1975). A case study in techniques of allocation. Biometrics, Vol. 31, 931-942.

(38) Garnett, James M. & Stephen S. Yau (1977). Nonparametric estimation of the bayes error of feature extractors using ordered nearest neighbor sets. IEEE Trans. on Computers, Vol. C-26, No. 1, 46-54.

(39) Gelperin, David (1977). On the optimality of A*. Artificial Intelligence, Vol. 8, 69-76.

(40) Gleason, C., R. Starbuck, R. Sigman, G. Hanuschak, M. Craig, P. Cook, & R. Allen (1977). The auxiliary use of landsat data in estimating crop acreages: results of the 1975 Illinois crop-acreage experiment. U.S. Department of Agriculture, Washington, D.C. Report SRS-21.

(41) Glick, Ned (1976). Sample-based classification procedures related to empiric distributions. IEEE Trans. on Information Theory, IT-22, No. 4, 454-461.

(42) Glick, Ned (1977). Additive estimators for probabilities of correst classification. Proceedings of the IEEE Computer Society Conference on Pattern Recognition and Image Processing, 304-317. IEEE Catalog # 77CH1208-9 C.

(43) Gnanadesikan, R. (1977). Statistical Data Analysis of Multivariate Observations. John Wiley.

(44) Goldstein, Matthew (1975). Comparison of some density estimate classification procedures. Journal of the American Statistical Assoc., Vol. 70, No. 351, 666-669.

(45) Goldstein, Matthew & Edward Wolf (1977). On the problem of bias in multinomial classification. Biometrics, Vol. 33, 325-331.

(46) Golomb and L.D. Baumert (1975). Backtrack programming. Journal ACM, Vol. 12, 516-524.

(47) Gonzalez, R.C. & L.C. Howington (1977). Machine Recognition of abnormal behavior in nuclear reactors. IEEE Transactions on Systems, Man and Cybernetics, Vol. SMC-7, No. 7, 717-728.

(48) Grenander, U. (1976). Pattern Synthesis. Lectures in Pattern Theory, Vol. 1, Springer-Verlag.

(49) Hall, P.A.V. (1973). Equivalence between and/or graphs and context free grammars. Communications ACM, Vol. 16, 444-445.

(50) Harris, L.R. (1974). The heuristic search under conditions of error. Artificial Intelligence, Vol. 5, 217-234.

(51) Hart, P.E., N.J. Nilsson, & B. Raphael (1968). A formal basis for the heuristic determination of minimal cost paths. IEEE Transactions on Systems Science Cybernetics, SSC-4, 100-107.

(52) Hauska, Hans & Philip H. Swain (1975). The decision tree classifier: design and potential. Proceedings of the 2nd Symposium on Machine Processing of Remotely Sensed Data, 38-48.

(53) Hoffman, W.C. (1976). An informal, historical description (with bibliography), of the Lie Transformation Group Approach to Neuropsychology. Department of Math., Oakland University, Rochester, Michigan.

(54) Horowitz, S.L., (1975). A syntactic algorithm for peak detection in waveforms with applications to cardiography, CACM, Vol. 18, No. 5.

(55) International Symposium on Information Theory, Abstract of Papers (1977), Cornell University, Ithica, N.Y.
IEEE Catalog # 77CH1277-3 IT.

(56) ICCS, Proceedings 1977 International Conference on Cybernetics and Society, Washington, D.C.
IEEE Catalog # 77CH1259-1 SMC.

(57) Jackson, Thomas J. (1976). The Value of Lindsat Data in Urban Water Resources Planning. Dissertation, University of Maryland Graduate School.

(58) Jain, Anil K. & Richard Dubes (1976). Feature definition in pattern recognition with small sample size. Michigan State University, College of Engineering, Department of Computer Science, Technical Report TR 76-04.

(59) Jelinek, Frederick (1976). Continuous speech recognition by statistical methods. Proceedings of the IEEE, Vol. 64, No. 4, 532-556.

(60) Kalensky, Z. & J.M. Wightmass (1976). Automatic Forest Mapping Using Remotely Sensed Data. paper presented at XVI IUFRO World Congress, Oslo; available from Forest Management Institute, Department of Environment, Ontario Canada.

(61) Kanal, L. (1974). Patterns in Pattern Recognition, 1968-1974. IEEE Transactions on Information Theory, Vol. IT-20, No. 6.

(62) Kanal, L. (1975). Prospects for Pattern Recognition. Electronics Industries Association, Washington, D.C.

(63) Kanal, L. (1977). Current status, problems, and prospects of pattern recognition. in Current Perspectives in Pattern Recognition-Special Issue, Systems, Man and Cybernetics Review, Newsletter of the IEEE SMC Society.

(64) Kanal, L.N. and A.V. Kulkarni (1976). Admissible strategies in a state space model for multistage multiclass classification. Abstracts, Conf. Proc. IEEE International Symposium Information Theory, Ronneby, Sweden.
IEEE Catalog # 76 CH0959IT.

(65) Kanal, L.N. & J.A. Parikh (1977). Severe Storm Pattern Recognition from Meteorological Satellite Data-A Report on Current Status and Prospects. Report # ECOM-77-3, Atmospheric Sciences Lab, U.S. Army Electronics Command, White Sands Missile Range, N.M.

(66) Knuth, D., (1971). Optimum binary search trees. Acta Informatica, Vol. 1, 14-25.

(67) Knuth, D. & R.W. Moore (1975). An analysis of Alpha-Beta pruning. Artificial Intelligence, Vol. 6. 293-326.

(68) Kohler, W.H., & K. Steiglitz (1974). Characterization and theoretical comparison of branch-and-bound algorithms for permutation problems. Journal ACM, Vol. 21, 140-156.

(69) Krzanowski, W.J. (1975). Discrimination and classification using both binary and continuous variables. Journal of the American Statistical Association, Vol. 70, No. 352, 782-790.

(70) Krzanowski, W.J. (1976). Canonical representation of the location model for discrimination or classification. Journal of the American Statistical Association, Vol. 71, No. 356, 845-848.

(71) Krzanowski, W.J. (1977). The performance of Fisher's linear discriminant function under non-optimal conditions. Technometrics, Vol. 19, No. 2, 191-200.

(72) Kulkarni, A.V. (1976). Optimal and heuristic synthesis of hierarchical classifiers. Ph.D. dissertation, University of Maryland, Computer Science Technical Report Series, TR-469.

(73) Kulkarni. A.V. & L.N. Kanal (1976). An optimization approach to hierarchical classifier design. Proc. 3rd International Joint Conference on Pattern Recognition, San Diego.

(73b) Kulkarni, A.V. & L.N. Kanal (1978), "Admissible search strategies for parametric and nonparametric hierarchical classifiers", Technical Report, Laboratory for Pattern Analysis, Department of Computer Science, University of Maryland.

(74) Ledley, R.S. (1966). Pattern recognition studies in the biometrical sciences. AFIPS Conference Proceedings SJCC, 411.

(75) Levi, G. & F. Sirovich (1976). Generalized and/or graphs. Artificial Intelligence, Vol. 7, 243-259.

(76) Liou, Jiunn-I and Richard Dubes (1977). A constructive method for grammatical inference based on clustering. Michigan State University, College of Engineering, Department of Computer Science, Technical Report TR 77-01.

(77) Lu, S.Y., & Fu, K.S. (1976). Efficient error-correcting syntax analysis for recognition of noisy patterns. School of Elec. Engineering, Purdue University, TR-EE 76-9, West Lafayette, Indiana.

(78) Lyons, G., (1974). Syntax-directed least-errors analysis for context-free languages: a practical approach, CACM , Vol. 17, No. 1, 3-13.

(79) MacLachlan, G.J. (1975). Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis. Journal of the American Statistical Association, Vol. 70, No. 350, 365-369.

(80) MacLachlan, G.J. (1976). The bias of the apparent error rate in discriminant analysis. Biometrika, Vol. 63.

(81) MacLachlan, G.J. (1977). A note on the choice of a weighting function to give an efficient method for estimating the Prob. of misclassification. Pattern Recognition, Vol. 9, 147-149.

(82) Martelli, Alberto (1977). On the complexity of admissible search algorithms. Artificial Intelligence, Vol. 8, 1-13.

(83) Mattson, R.L., Dammann, J.E. (1965). A technique for detecting and coding subclasses in pattern recognition problems. IBM Journal.

(84) Meisel, W.S., Michalopoulos, D.S. (1973). A partitioning algorithm with application in pattern classification and the optimization of decision trees. IEEE Transactions on Computers, Vol. C-22, 93-103.

(85) Miller, P.L. (1973). A locally organized parser for spoken output. Tech. Report 503, Lincoln Lab. M.I.T., Cambridge Mass.

(86) Moore, David S., S. J. Whitsitt & David Landgrebe (1976). Variance comparisons for unbiased estimators of probability of correct classification. IEEE Trans. on Information Theory, 102-105.

(87) Morgera, S.D. & David B. Cooper (1976). On the role of dimensionality and sample size for unstructured and structured covariance matrix estimation. Third International Joint Conference on Pattern Recognition, 467-472.

(88) Nadler, M. (1971). Error and Reject Rates in a Hierarchical Pattern Recognizer. IEEE Trans. Comp. Vol. C-20.

(89) Nilsson, N.J. (1971). Problem-Solving Methods in Artificial Intelligence. McGraw-Hill, New York.

(90) Parikh, JoAnn (1977). A comparative study of cloud classification techniques. Remote Sensing of Environment, Vol. 6, 67-81.

(91) Patterson, C.L.(1977). Special Issue on Image Processing, Computer, IEEE Computer Society, Vol. 10, No. 8.

(92) Pavlidis, T. (1976). Syntactic pattern recognition on the basis of functional approximation. Pattern Recognition and Artificial Intelligence, Academic Press.

(93) Pavlidis, T. (1977). Structural Pattern Recognition. Springer-Verlag.

(94) Pollack, S.L. (1971). Decision Tables-Theory and Practice. Wiley Interscience, New York.

(95) Pressman, Norman J. (1976). Markovian analysis of cervical cell images. Journal of Histochemistry and Cytochemistry, Vol. 24, No. 1, 138-144.

(96) Proceedings, The Third International Joint Conference on Pattern Recognition, November 8-11,(1976), Coronada, Calif. IEEE Catalog # 76 CH140-3C.

(97) Proceedings of the Workshop on Pattern Recognition Applied to Oil Identification (1976), Coronada, Calif. IEEE Catalog # 76 CH1247-6C.

(98) Proceedings of the Symposium on Computer-Aided Diagnosis of Medical Images, (1976), Nov. 11. IEEE Catalog #76CH1170-OC.

(99) Proceedings of the 2nd Symposium on Machine Processing of Remotely Sensed Data. (1975), Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana. IEEE Catalog #75 CH1009-0-C.

(100) Proceedings of the International Symposium on Computer Aided Seismic Analysis and Discrimination, (19), Falmouth, Mass. IEEE Cat. No. 77CH1244-3C.

(101) Ragan, Robert M., Thomas J. Jackson, Richard H. McCuen, Carl D. Ealy, & Mark L. Mercer (1976). Use of Landsat and high altitude aircraft remote sensing in urban hydrology. University of Maryland, College Park, Civil Engineering Department.

(102) Reason, C.C. (1976). A bi-directional speech parsing technique. TR-90, University of Toronto, Department of Computer Science.

(103) Reddy, D.R., Erman, L.D., Fennell, R.D., and Neely, R.B. (1973). The HEARSAY speech understanding system. Proc. 3rd International Joint Conf. on Artificial Intelligence, 185-193.

(104) Reddy, R. (Ed.) (1975). Speech Recognition. Academic Press.

(105) Reinwald, L.T., Soland, R.M.(1966). Conversion of limited entry decision tables to optimal computer programs I-Minimum average processing time. Journal ACM, Vol. 13, 339.

(106) Reinwald, L.T., Soland. R.M. (1967). Conversion of limited entry decision tables to optimal computer programs II-Minimum storage requirements. Journal ACM, Vol. 14, 742.

(107) Rosenfeld, A. (Ed), (1976). Digital Picture Analysis. Springer-Verlag.

(108) Stallings, William (1977). Fuzzy set theory versus Bayesian statistics. IEEE Trans. on SMC, Vol. 9, 216-222.

(109) Starks, Scott A. and R.P. deFigereido (1977). A new approach to structure preserving feature extraction. Preprint, Proc. 1977 Conference on Information Sciences and Systems,Johns Hopkins University, Baltimore, Maryland.

(110) Stockman, George (1977). A problem-reduction approach to the linguistic analysis of waveforms. Ph.D. Dissertation. U. of Maryland, Computer Science Technical Report TR-538.

(111) Stockman, George (1977). Non-directional pattern recognition using a problem reduction representation. *Proc. 1977 International Conference on Cybernetics and Society*, 601-604.

IEEE Catalog # 77 CH1259-ISMC.

(112) Stockman, G., L. Kanal, and M.C. Kyle (1976). Structural pattern recognition of carotid pulse waves using a general waveform parsing system. *Comm. ACM*. Vol. 19, No. 12, 688-695.

(113) Teebe, D.L. (1976). A nearest neighbors spectrum of the Bayes Risk Function. *Proceedings 3rd Joint International Conference on Pattern Recognition*, 253-254.

IEEE Catalog # 76 CH1140-3C.

(114) Toussaint. G.T. & Sharpe, P.M. (1973). An efficient method for estimating the probability of misclassification applied to a problem in medical diagnosis. School of Computer Science, McGill Univ. Montreal, Canada.

(115) VanderBrug, Gordon, J. (1976). Problem Representations and Formal Properties of Heuristic Search. *Information Sciences*, Vol. 11, 279-307.

(116) VanderBrug, G.J. & J. Minker (1975). State-space, problem reduction and theorem-proving-some relationships. *Communications ACM*. Vol. 18, 107-115.

(117) Van Ness, J.W. (1977). Dimensionality and classification performance with independent coordinates. *IEEE Trans. on Systems Man & Cybernetics*, Vol SMC-7, No. 7, 560-654.

(118) Wald, A. (1947). *Sequential Analysis*. Wiley, New York.

(119) Walker, D.E. (1974). The SRI speech understanding system. *IEEE Symposium on Speech Recognition*. L. Erman (ed).,

(120) Waller, W.G. and Anil K. Jain (1976). Mean Recognition Accuracy of dependent binary measurements. Michigan State University, Department of Computer Science Technical Report TR-76-03.

(121) Whitsitt, S.J. and D.A. Landgrebe (1977). Error estimation and separability measures in feature selection for multiclass pattern recognition. Lab for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana.

(122) Williams, W.J. (ed), (1977). *Proceedings* of the IEEE, Special Issue on Biological signal processing and analysis.

(123) Winston, P. (1969). A heuristic program that constructs decision trees. MIT Project MAC, Memo 173.

(124) Winston, P.H. (1977) Artificial Intelligence. Addison-Wesley.

(125) Wold, Svante (1976). Pattern recognition by means of disjoint principal components models. Pattern Recognition, Vol. 8, 127-139.

(126) Wu, Chialin, D.Landrebe, & P. Swain (1975). The decision tree approach to classification. Purdue University, School of Electrical Engineering, Report TR-EE 75-17.

(127) You, K.C. and K.S. Fu (1976). An approach to the design of a linear binary tree classifier. Proc. 3rd Symp. Machine Processing of Remotely Sensed Data, LARS, Purdue University.

(128) Yunck, Thomas P. (1976). A technique to identify nearest neighbors. IEEE Trans. on SMC. Vol. SMC-6, No. 10, 678-683.

Keywords: Pattern Recognition, Review, Multistage Classification, Structural Feature Extraction, Nearest Neighbor, State-Space, AND/OR graph, Ordered Search.

ABSTRACT

Noting the major limitations of the much developed multivariate statistical and syntactic pattern recognition models, this paper describes in a tutorial manner alternate representations, based on stage-space and AND/OR graphs and ordered search strategies, for multistage and nearest neighbor classification and for structural pattern analysis and feature extraction. Some recent work in pattern recognition is reviewed from these vantage points.

In addition, the paper touches on recent contributions to the continuing attempts to understand feature subset selection, measurement complexity and nonparametric classification and error estimation. Surveys, conference proceedings and edited collections providing quick access to the recent literature on pattern recognition methodologies and applications, are cited in the bibliography.

20.

In addition, the paper touches on recent contributions to the continuing attempts to understand feature subset selection, measurement complexity and nonparametric classification and error estimation. Surveys, conference proceedings and edited collections providing quick access to the recent literature on pattern recognition methodologies and applications, are cited in the bibliography.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFOSR-TR- 78-0958	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) SOME CURRENT CONCEPTS AND PROBLEMS IN PATTERN CLASSIFICATION AND FEATURE EXTRACTION		5. TYPE OF REPORT & PERIOD COVERED Interim 6. PERFORMING ORG REPORT NUMBER
7. AUTHOR(s) Laveen N. Kanal		8. CONTRACT OR GRANT NUMBER(s) AFOSR 76-2901 new
9. PERFORMING ORGANIZATION NAME AND ADDRESS Laboratory for Pattern Analysis University of Maryland College Park, MD 20742		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61102F 2304/A2
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Office of Scientific Research/NM Bolling AFB, Washington, D.C. 20332		12. REPORT DATE April 1978 13. NUMBER OF PAGES 34
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Pattern Recognition, Review, Multistage Classification, Structural Feature Extraction, Nearest Neighbor, State-Space, AND/OR graph, Ordered Search.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Noting the major limitations of the much developed multi-variate statistical and syntactic pattern recognition models, this paper describes -- in a tutorial manner -- alternate representations, based on stage-space and AND/OR graphs and ordered search strategies, for multistage and nearest neighbor classification and for structural pattern analysis and feature extraction. Some recent work in pattern recognition is reviewed from these vantage points. → (over)